https://doi.org/10.1016/j.patcog.2019.04.028

Deep-Feature Encoding-based Discriminative Model for Age-invariant Face Recognition

M. Saad Shakeel* and Kin-Man Lam

Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong

ABSTRACT

Facial aging variation is a major problem for face recognition systems due to large intra-personal variations caused by age progression. A major challenge is to develop an efficient, discriminative feature representation and matching framework, which is robust to facial aging variations. In this paper, we propose a robust deep-feature encoding-based discriminative model for age-invariant face recognition. Our method learns high-level deep features using a pre-trained deep-CNN model. These features are then encoded by learning a codebook, which converts each of the features into a discriminant *S*-dimensional codeword for image representation. By incorporating the locality information in the whole learning process, a closed-form solution is obtained for both the codebook-updating and encoding stages. As the features of the same person at different ages should have certain correlations, canonical correlation analysis is utilized to fuse the pair of training features, for two different ages, to make the codebook discriminative in terms of age progression. In the testing stage, the gallery and query image's features are encoded using the learned codebook. Then, linear mapping based on linear regression is employed for face matching. We evaluate our method on three publicly available challenging facial aging datasets, FGNET, MORPH Album 2, and Large Age-Gap (LAG). Experimental results show that our proposed method outperforms various state-of-the-art age-invariant face recognition methods, in terms of the rank-1 recognition accuracy.

Keywords—Age-invariant face recognition, canonical correlation analysis, deep learning, discriminative model, feature encoding, linear regression.

1. Introduction

Face recognition in unconstrained environments is one of the most widely studied topics in the pattern recognition community, and has achieved remarkable progress in recent years due to better feature extraction and matching frameworks. Despite the progress, recognizing face images with large age-variations require considerable amount of attention. Age-invariant face recognition (AIFR) has many practical applications, e.g. criminal identification using

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

photographs, finding missing children, etc. The major challenge is the large intra-personal variations due to the complex age progression, as shown in Fig.1. Age progression is a non-linear process, which greatly affects the facial appearance in two phases of life. From childhood to adulthood, the greatest change occurs in terms of shape, while change in texture takes place from adulthood to old age. All these factors create hurdles in learning age-invariant patterns.

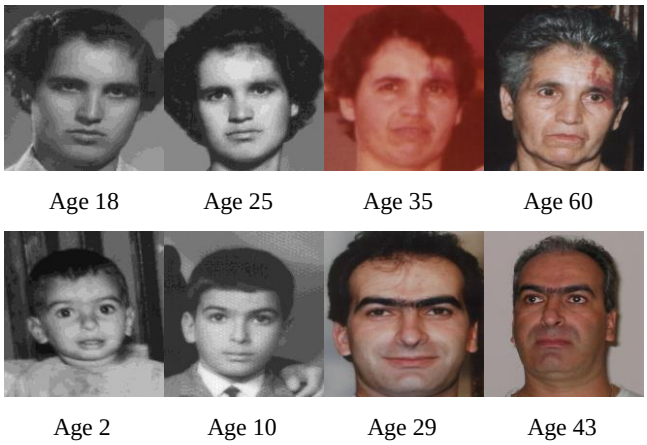


Fig.1 Sample images from the FGNET data set from two different persons with large age variations, where each row represents the face images of the same person.

The most important part of a face recognition system is to extract distinctive features from face images to reduce the intra-personal variations, caused by illumination, expression, pose, age, etc. Among the existing feature descriptors, local features have shown high robustness to pose, expression, and illumination variations. However, these features are not optimal for solving AIFR problem [1] and provide limited performance. Furthermore, their performance heavily depends on the properly preprocessed face images. To make these features more discriminative for recognition, various feature-encoding-based methods [2-8] have been proposed, which convert extracted features into a discriminative N-dimensional codeword for image representation. This brings major improvements in the recognition performance under unconstrained environments. Recently, deep-learning methods [9-12] have also gained a lot of attention in the face-recognition community, due to their superior performance. Those methods have already achieved more than 99% recognition rate on the challenging Labeled Faces in the Wild (LFW) data set, which even surpasses the human-level performance. However, their performance is limited in solving the aging face-recognition problem.

To tackle the abovementioned challenges, we propose a robust feature-encoding method based on locality constraint, which converts extracted deep features into an S-dimensional codeword for face representation. In this regard, we first learn an age-discriminative codebook, which uses the same codeword to represent the same identity at different ages.

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

Our method imposes locality constraint in the whole learning process, which generates an optimized codebook that better fits the data's local structure. Furthermore, it reduces the intra-personal variations by sharing the same local bases of the learned codebook for images of the same identity. However, dissimilar bases are shared among the images of different identities, so this enhances the inter-personal variations. Using a pair of face images with a large age difference, we first exploit their correlation by projecting them into a coherent feature subspace using Canonical Correlation Analysis (CCA), and then perform feature fusion. Those fused features are then used to learn an age-discriminative codebook. In the testing stage, the gallery and query image's features are encoded using the learned codebook. To accelerate the encoding process, only a few codebook entries, which are close to the query feature, are selected for reconstruction. The selection is based on the k-NN search strategy. Furthermore, the resultant encoded features are sparse, which enhances the discriminative power of the learned features. In the final stage, the linear-regression model [13] is employed to determine the relationship between the encoded gallery and the query image's features at different ages, in terms of a coefficient vector. By using this coefficient vector, residual values are computed for face matching.

1.1 Motivation

Data locality is considered as a key issue in numerous computer vision applications, such as dimensionality reduction, clustering, image classification, etc. Wang et al. [8] proposed a feature-encoding framework based on locality constraint, namely Locality Constrained Linear coding (LLC), which projects the extracted features into a local coordinate system. It was argued that locality is more essential than sparsity, as locality will always lead to sparsity for the resultant encoded coefficients, but not vice versa. Motivated by this idea, we learn an age-discriminative codebook by taking the data's local structures into account. One of the major advantages of locality constraint is that it captures the correlation between the features of the same identity by sharing the local bases of the codebook. Therefore, this ensures that the same codeword is used to represent face images of the same identity taken at different ages. For cross-age face recognition, these codewords are explored.

The major contributions of this paper are as follows:

- A robust feature-encoding framework based on locality constraint is proposed, which encodes the extracted deep facial features into a discriminative codeword. In comparison to LLC [8], our algorithm incorporates the locality

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

information during the whole learning process and provides closed form solutions for both sparse coding and the codebook updating stage.

- In the training stage, we maximize the correlation between the deep features of the same identity with a large age difference using CCA, which are then used to learn an age-discriminative codebook. The learned codebook is proved to be robust to age variations, as supported by our experimental results.
- Extensive experiments have been conducted on three challenging face-aging data sets, and experiment results show that our proposed method outperforms other state-of-the-art age-invariant face recognition methods, in terms of recognition rate. Furthermore, our proposed method shows robustness to externally added noise, and achieves superior performance, as verified by our experimental results.

The rest of the paper is organized as follows. In Section 2, we briefly review some related work of age-invariant face recognition. In Section 3, we introduce the process of deep-feature extraction using a pre-trained deep-CNN model (AlexNet). In Section 4, we present our feature-encoding framework based on the Euclidean locality adaptor, which makes the extracted deep features more robust to age variations. In Section 5, we explain the concept of feature fusion using CCA. In Section 6, we discuss the use of linear regression for classification. Section 7 presents the experimental results of our proposed discriminative model, along with discussions. Finally, Section 8 concludes our paper.

2. RELATED WORK

Existing research work, related to facial aging, mainly focuses on either age estimation [14-22] or age simulation [23-30]. A typical approach to age-invariant face recognition is to synthesize a test face image to be the same age as the gallery image before performing recognition, i.e. it is based on generative models. These generative model-based methods have some limitations. First, constructing an accurate aging model requires strong parametric assumptions, as well as the real ages of the training images, which makes it unsuitable for real-time face recognition systems. Furthermore, it divides the whole recognition framework into two steps, which creates difficulties in achieving optimal recognition performance in an end-to-end manner. Moreover, the synthesized face images bring some additional noises, which has a severe impact on recognition performance.

Recently, discriminative models have been proposed for AIFR, which focus on developing features insensitive to facial aging. Ling et al. [31] utilized the Gradient Orientation Pyramid (GOP) for feature extraction, and then used

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

Support Vector Machine (SVM) for classification. In [32], the feature descriptors, Multi-scale Local Binary Pattern (MLBP) and Scale Invariant Feature Transform (SIFT) were utilized, and a fusion framework, based on random sampling, was proposed to improve the recognition performance. Gong et al. [33] proposed a probabilistic model, which represents a face image as a combination of two components, identity and age, that affect facial appearance. A learning process based on the Expectation-Maximization (EM) algorithm was proposed, which estimates aging and identity features simultaneously. Later, a novel feature descriptor [34], namely Maximum Entropy Feature Descriptor (MEFA), was proposed, which analyses facial images on a micro level and encodes the information in a form of discrete codes. The approach is based on a new matching framework, known as identity factor analysis, which estimates the probability that the two given face images are from the same identity. Chen et al. [35] proposed a new coding framework, which performs feature encoding by using reference images with an age-invariant reference space. Li et al. [36] proposed a hierarchical model with two stages of learning. Firstly, discriminative features are learned from microstructures of facial images, which are then converted into integer codes for face recognition. Recently, Zhou et al. [37] proposed an identity-inference model, which uses appearance-age labels to learn the aging subspace. The model utilizes probabilistic linear discriminant analysis (PLDA) to represent the identity and aging components. After that, the identity subspace is determined in an iterative way using the EM algorithm. The learned identity component is then used to perform face recognition.

2.1 Deep-learning-based Methods

Deep-learning-based models [38-41] handles age-variations by learning age-invariant face representation with specific loss functions. Inspired by this, Wen et al. [38] learned age-invariant deep features in an end-to-end manner based on latent identity analysis. The method keeps the identity components separated from the aging variations in learning the deep features. However, it assumes that the combination of the identity and aging features is linear. Hence, it ignores the correlation between the aging and identity component. Another deep learning framework [39] for age-invariant face recognition was proposed, which shares the same feature layers with the identity and the aging networks. Both networks are trained alternately to separate the identity features from the aging features. Xu et al. [40] proposed an auto-encoder network to learn the complex non-linear aging progression. The method decomposes a face image into identity, age, and noise components by proposing a non-linear factor analysis method. Recently, a novel distance metric

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

optimization method [41] based on deep learning was proposed. The method used large number of matched pairs from the training set to learn the identification information. These matched pairs then serve as an input to enhance the differences between unmatched pairs. The model parameters are updated using the classical gradient descent algorithm. Therefore, the learned features and distance metric are optimized simultaneously. However, the major drawbacks of deep learning-based methods are the high computational cost and the large amount of labeled training data.

2.2 Feature-encoding techniques

Bag of Features (BOF) [2] is one of the earliest proposed methods, which represents the extracted local features in terms of a histogram, based on a set of learned vocabularies. The major drawback of this method is that it does not consider the layout structure of the features, which makes it unsuitable for capturing the shape of an object. Many extensions [3-5] of BOF were proposed, which can be classified into two categories, the generative model and the discriminative model based on codebook learning. These methods have achieved superior performances by using Spatial Pyramid Matching (SPM). This matching approach first computes feature descriptors from densely located feature points, and then applies the learned vocabularies or codebook with N entries to convert the descriptors into an N-dimensional codeword. To further increase the scalability, Yang et al. [6] proposed using sparse coding to obtain non-linear codes for non-linear feature representations. Yu et al. [7] achieved an improvement to the Sparse Coding (SC)-based approach by proposing a model, namely Local Coordinate Coding (LCC), which performs feature-encoding based on the locality information. Like SC, the model also needs to solve the l_1-norm minimization problem, which makes it computationally expensive. A fast implementation of LCC was proposed in [8], which projects each feature descriptor into its local coordinate system using locality constraint. This approach preserves the local information, which favors both better feature representation and classification.

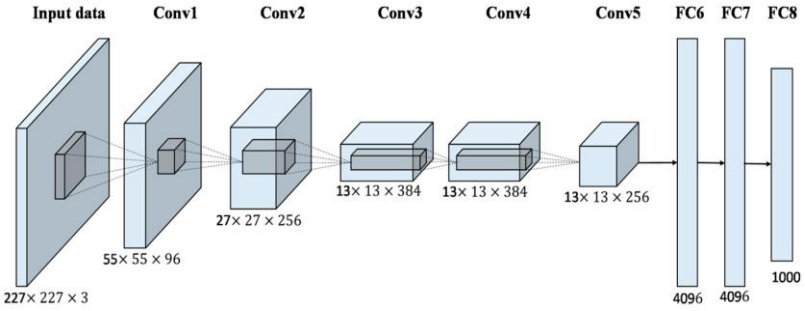


Fig.2. AlexNet architecture (Adapted from [62]).

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

3. Discriminative Model

Our proposed discriminative model consists of four main parts: (1) deep-feature extraction, (2) age-discriminative codebook learning, (3) feature-encoding using learned codebook, and (4) face matching based on linear mapping. We will explain all these stages in the following sections.

3.1 Deep-Feature Extraction

In our work, we utilize a pre-trained deep-CNN model, namely AlexNet [42], to extract high-level deep features. AlexNet is selected, due to its simple architecture and superior performances. Fig. 2 shows the architecture of AlexNet. CNN models have the capability of extracting both low-level, as well as high-level, features from input images. AlexNet consists of 5 convolutional layers, 3 pooling layers, and 3 fully connected (FC) layers. The output of each convolutional and FC layer is fed to the ReLU activation function. The first convolutional layer filters the input face image of size $227 \times 227 \times 3$, with 96 kernels of size $11 \times 11 \times 3$, using a stride of 4 pixels. The output of the first convolutional layer is fed to the second convolutional layer, after passing through the normalization and pooling layer. Finally, the network learns high-level deep features at the last FC layer, with a dimension of 4096. All fully connected layers are regularized by using a drop-out scheme. In our method, we extract features from the FC layer (fc7). In deep learning-based methods, networks are trained for feature extraction and recognition from end to end. In our proposed approach, we use deep-learning-based CNN model to extract features, which are then converted into discriminative codewords for recognition. Fig.3 shows the features learned at the different convolutional layers of the CNN.

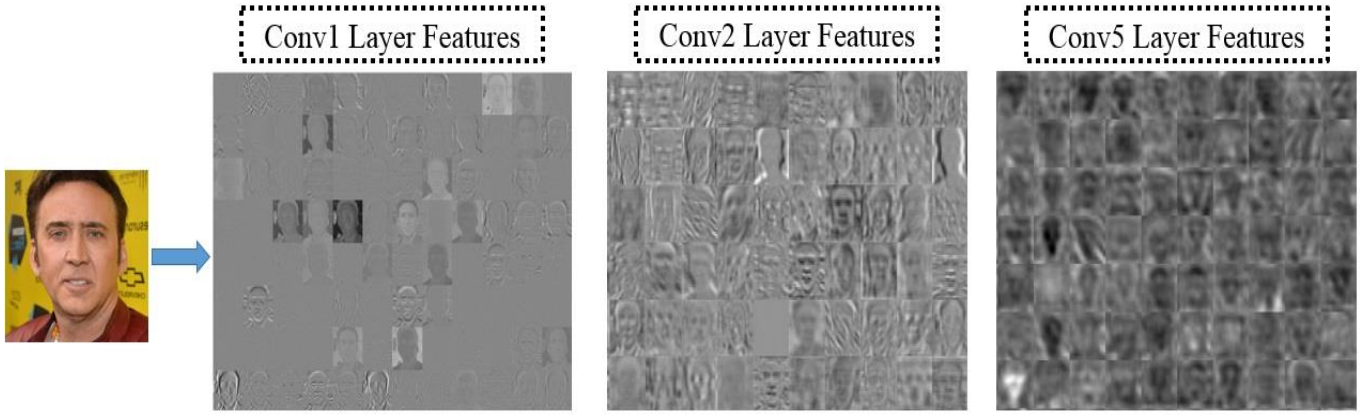


Fig.3. Visualization of the learned deep features from different convolutional layers of AlexNet.

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

4. Feature Encoding based on Locality Information

We now explain our proposed feature-encoding framework, based on the Euclidean locality adaptor. Before we proceed, we first explain the major differences between data sparsity and data locality. The advantages of using locality information for accurate classification (recognition) will also be discussed. Our proposed method provides a better feature-encoded representation, which enhances the discriminability of the extracted deep features in terms of age progression.

4.1 Locality vs Sparsity

Data locality information has proven to be important for the success of various pattern-recognition applications, e.g. density estimation [43], dimensionality reduction [44], and image classification [45]. Sparse Representation-based Classification (SRC) was used for face recognition under large occlusion and achieved state-of-the-art performances. However, is it enough, or essential, for resolving this kind of a problem? This query has recently been examined in [8], which concluded that exploitation of sparse information only is not enough to handle pattern recognition with large occlusion.

As discussed before, data locality is more necessary than sparsity, as locality will also lead to sparsity for the resultant encoded coefficients, but not vice versa. By using locality information, codebook entries near the query input will be selected for the reconstruction of data samples. However, classification based on sparse representation minimizes the class-wise reconstruction error, so it may use codebook entries far away from the query input for reconstruction, and this is not desirable. According to the assumption made by the *k*NN classifier, those codebook entries far away from the input have less probability of belonging to the same class. The standard feature-encoding method [6], based on data sparsity, does not preserve the local structure of data samples during the encoding stage, but LLC [8] does. Like locality constraint, sparse coding also leads to a low reconstruction error by utilizing number of local bases from a learned codebook. However, the regularization term used in SC is not smooth. Furthermore, SC can select different bases for the similar data samples, which results in a loss of correlation between the codewords. The use of locality constraint ensures that the same or similar codewords for the images belonging to the same class will be used. Moreover, the optimization techniques required by SC are computationally expensive. This motivates us to propose a feature-

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

encoding-based framework based on the locality constraint, which preserves the data structure and can achieve a better feature representation, and hence, enhances the recognition performance.

4.2 Locality-based feature-encoding framework

In this section, we present our feature-encoding framework based on the Euclidean locality adaptor to achieve a more compact and discriminative feature representation. We propose to encode the features by projecting each of them into its local coordinate system. Consider that M feature vectors of dimension D are extracted, which are represented as $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \dots, \boldsymbol{f}_M] \in R^{D \times M}$. Firstly, a codebook $\boldsymbol{W}$ with S entries, i.e. $\boldsymbol{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_S\}$, is generated using k-means clustering, which is then used to convert each local descriptor into a S-dimensional codeword for final image description. Our proposed feature-encoding approach focuses on the Euclidean locality constraint rather than the sparsity constraint. Our proposed locality-based objective function is defined as follows:

$$\min_{\boldsymbol{W},\boldsymbol{C}} \|\boldsymbol{F} - \boldsymbol{W}\boldsymbol{C}\|^2 + \lambda \sum_{k=1}^{M} \|\boldsymbol{l}_k \otimes \boldsymbol{c}_k\|_2^2, \quad \text{s.t.}\ \mathbf{1}^T \boldsymbol{c}_k = 1, \tag{1}$$

where $\otimes$ represents the element-wise multiplication operator, λ is a regularization parameter, $\boldsymbol{l}_k \in R^S$ represents the locality term that consists of a Euclidean adaptor, which provides freedom to each basis vector depending on its similarity to the given descriptor $\boldsymbol{f}_k$. $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \dots, \boldsymbol{c}_M]$ is the set of codes for $\boldsymbol{F}$.

Each entry of the locality term $\boldsymbol{l}_k$ can be defined as follows:

$$\boldsymbol{l}_{ks} = (\sigma^2 + \|\boldsymbol{f}_k - \boldsymbol{w}_s\|^2)^{-1}\ , \tag{2}$$

where $\|\boldsymbol{f}_k - \boldsymbol{w}_s\|^2$ is the Euclidean distance between the deep-feature $\boldsymbol{f}_k$ and the s^{th} codebook entry $\boldsymbol{w}_s$, and σ is a constant used to control the weight decay speed for the locality term. In our method, we choose $\sigma = 0.5$. The locality term in equation (2) is the Euclidean adaptor, which defines how local coding varies with respect to the distances $\|\boldsymbol{f}_k - \boldsymbol{w}_s\|$. It utilizes the student t-distribution to provide a degree of freedom, such as Cauchy distribution. One of the major properties of the student t-distribution is that $(\sigma^2 + \|\boldsymbol{f}_k - \boldsymbol{w}_s\|^2)^{-1}$ follows an inverse square law when the pairwise distance $\|\boldsymbol{f}_k - \boldsymbol{w}_s\|^2$ is large. $\boldsymbol{l}_k$ is further normalized to have a value between 0 and 1 by taking a difference between $\max(\|\boldsymbol{f}_k - \boldsymbol{w}_s\|^2)$ and $\|\boldsymbol{f}_k - \boldsymbol{w}_s\|^2$. Local bases $\boldsymbol{w}_k$ are selected for each feature descriptor, such that a local

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

coordinate system can be built. Local bases can be considered as the nearest neighbors of $\boldsymbol{f}_k$, which lead to a more compact and simplified linear system for feature coding.

To solve (1), we utilize the Lagrange multiplier, which is defined as follows:

$$L(\boldsymbol{c}_k, \eta) = \|\boldsymbol{f}_k - \boldsymbol{W}\boldsymbol{c}_k\|^2 + \lambda\|\boldsymbol{l}_k \otimes \boldsymbol{c}_k\|_2^2 + \eta(\mathbf{1}^T\boldsymbol{c}_k - 1). \tag{3}$$

Let $\boldsymbol{Y} = (\boldsymbol{f}_k\mathbf{1}^T - \boldsymbol{W})^T(\boldsymbol{f}_k\mathbf{1}^T - \boldsymbol{W}) \in R^{S\times S}$, which is symmetrical. Equation (3) can be rewritten as:

$$L(\boldsymbol{c}_k, \eta) = \boldsymbol{c}_k{}^T\boldsymbol{Y}\boldsymbol{c}_k + \lambda\boldsymbol{c}_k{}^T\mathrm{diag}(\boldsymbol{l}_k)^2\boldsymbol{c}_k + \eta(\mathbf{1}^T\boldsymbol{c}_k - 1)\,, \tag{4}$$

where $\mathrm{diag}(\boldsymbol{l}_k)$ is a diagonal matrix. To determine the optimal solution of (4), its partial derivative is set to zero, which gives the following equation:

$$\frac{\partial(L(\boldsymbol{c}_k,\eta))}{\partial(\boldsymbol{c}_k)} = 2\boldsymbol{Y}\boldsymbol{c}_k + 2\lambda\mathrm{diag}(\boldsymbol{l}_k)^2\boldsymbol{c}_k + \eta^T\mathbf{1} = 0\,. \tag{5}$$

Let $\boldsymbol{\Phi} = 2(\boldsymbol{Y} + \lambda\mathrm{diag}(\boldsymbol{l}_k)^2)$, we have

$$\boldsymbol{\Phi}\boldsymbol{c}_k + \eta\mathbf{1} = 0\ , \tag{6}$$

Multiply (6) by $\mathbf{1}^T\Phi^{-1}$, the following equation is obtained:

$$\mathbf{1}^T\boldsymbol{\Phi}^{-1}\boldsymbol{\Phi}\boldsymbol{c}_k + \eta(\mathbf{1}^T\boldsymbol{\Phi}^{-1}\mathbf{1}) = 0 \tag{7}$$

According to the constraint $\mathbf{1}^T\boldsymbol{c}_k = 1$, $\mathbf{1}^T\boldsymbol{\Phi}^{-1}\Phi\boldsymbol{c}_k = 1$, (7) becomes:

$$1 + \eta(\mathbf{1}^T\boldsymbol{\Phi}^{-1}\mathbf{1}) = 0,$$

$$\eta = -(\mathbf{1}^T\boldsymbol{\Phi}^{-1}\mathbf{1})^{-1}. \tag{8}$$

Putting η in (8) into (6), we obtain the following equation:

$$\boldsymbol{\Phi}\boldsymbol{c}_k = (\mathbf{1}^T\boldsymbol{\Phi}^{-1}\mathbf{1})^{-1}\mathbf{1}.$$

After some transformations, we have

$$\boldsymbol{c}_k = \frac{\boldsymbol{\Phi}^{-1}\mathbf{1}}{\mathbf{1}^T\boldsymbol{\Phi}^{-1}\mathbf{1}} = \frac{\frac{1}{2}\boldsymbol{\Phi}^{-1}\mathbf{1}}{\mathbf{1}^T\left(\frac{1}{2}\boldsymbol{\Phi}^{-1}\mathbf{1}\right)}. \tag{9}$$

In this way, we obtain the analytical solution of our proposed objective function, which is given as follows:

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

$$\tilde{\boldsymbol{c}}_k = \frac{1}{2}\boldsymbol{\Phi}^{-1}\mathbf{1} = (\boldsymbol{Y} + \lambda \mathrm{diag}(\boldsymbol{l}_k)^2)^{-1}\mathbf{1},$$

$$\boldsymbol{c}_k = \tilde{\boldsymbol{c}}_k / (\mathbf{1}^T \tilde{\boldsymbol{c}}_k). \tag{10}$$

To update the codebook $\boldsymbol{W}$, it needs to solve the following equation:

$$\min_{\boldsymbol{W}} \|\boldsymbol{F} - \boldsymbol{W}\boldsymbol{C}\|^2 + \lambda \sum_{k=1}^{M} \|\boldsymbol{l}_k \otimes \boldsymbol{c}_k\|_2^2. \tag{11}$$

Let the objective function (11) be denoted as $\boldsymbol{\mathcal{F}}(\boldsymbol{W})$. Analytical solution of (11) can be derived by computing the partial derivative of $\boldsymbol{\mathcal{F}}(\boldsymbol{W})$ with respect to $\boldsymbol{w}_s$ for $s \in \{1,2,\dots,S\}$, which gives us the following equation:

$$\frac{\partial \boldsymbol{\mathcal{F}}}{\partial \boldsymbol{w}_s} = \sum_{k=1}^{M} -2c_{ks}(\boldsymbol{f}_k - \boldsymbol{W}\boldsymbol{c}_k) - 2\lambda c_{ks}^2(\boldsymbol{f}_k - \boldsymbol{w}_s) \tag{12}$$

In an equivalent form, it can be written as:

$$\left(\frac{\partial \boldsymbol{\mathcal{F}}}{\partial \boldsymbol{w}_s}\right)^T = \sum_{k=1}^{M} (-2c_{ks}(1 + \lambda c_{ks})(\boldsymbol{f}_k)^T + 2(\lambda c_{ks}^2 \boldsymbol{w}_s{}^T + c_{ks} \sum_{j=1}^{S} c_{kj} \boldsymbol{w}_j{}^T)) \,. \tag{13}$$

To compute the global minimum of (11), we set the partial derivatives in (13) to zero. After setting the partial derivative of (11) to zero for $s = 1, 2, 3, \dots, S$, we obtain

$$\boldsymbol{P}\boldsymbol{W}^T = \boldsymbol{Q}, \tag{14}$$

where the matrices $\boldsymbol{P} \in \mathbb{R}^{S \times S}$ and $\boldsymbol{Q} \in \mathbb{R}^{S \times d}$ are

$$\boldsymbol{P} = \sum_{k=1}^{M} \begin{pmatrix} (1+\lambda)c^2{}_{k1} & c_{k1}c_{k2} & \cdots & c_{k1}c_{kS} \\ c_{k1}c_{k2} & (1+\lambda)c^2{}_{k2} & \cdots & c_{k2}c_{kS} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1}c_{kS} & c_{k2}c_{kS} & \cdots & (1+\lambda)c_{kS}^2 \end{pmatrix},$$

$$\boldsymbol{Q} = \sum_{k=1}^{M} \begin{pmatrix} c_{k1}(1+\lambda c_{k1})(\boldsymbol{f}_k)^T \\ c_{k2}(1+\lambda c_{k2})(\boldsymbol{f}_k)^T \\ \vdots \\ c_{kS}(1+\lambda c_{kS})(\boldsymbol{f}_k)^T \end{pmatrix}. \tag{15}$$

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

Algorithm 1: Codebook updating scheme (Training stage)

Input: $\mathbf{W}_{init} \in R^{D\times S}, \boldsymbol{F} \in R^{D\times M}, \sigma, \lambda$
Output: **W**

1: $\mathbf{W} \leftarrow \mathbf{W}_{init}$ (Initialize the codebook using *k*-means)
2: **for** $k = 1{:}\,M$ **do**
3: $\boldsymbol{l} \leftarrow 1 \times S$ (Locality constraint)
4: **for** $k = 1{:}\,S$ **do**
5: $\boldsymbol{l}_{ks} \leftarrow (\sigma^2 + \|\boldsymbol{f}_k - \boldsymbol{w}_s\|^2)^{-1}$
6: end **for**
7: $\boldsymbol{l} \rightarrow normalize_{(0,1)}(\boldsymbol{l})$
8: Lagrange function to solve (1)

$L(\boldsymbol{c}_k, \eta) = \|\boldsymbol{f}_k - \boldsymbol{W}c_k\|^2 + \lambda\|\boldsymbol{l}_k \otimes \boldsymbol{c}_k\|_2^2 + \eta(\mathbf{1}^T\boldsymbol{c}_k - 1)$

$\boldsymbol{c}_k = \tilde{\boldsymbol{c}}_k/(\mathbf{1}^T\tilde{\boldsymbol{c}}_k)$ Eqs. (3) − (10) (Analytical solution)

9: Codebook Updating:

$$\left(\frac{\partial \boldsymbol{\mathcal{F}}}{\partial \boldsymbol{w}_s}\right)^T = \sum_{k=1}^{M}(-2c_{ks}(1+\lambda c_{ks})(\boldsymbol{f}_k)^T + 2(\lambda c_{ks}^2 \boldsymbol{w}_s{}^T + c_{ks}\sum_{j=1}^{S} c_{kj}\boldsymbol{w}_j{}^T))$$

10: $\boldsymbol{P}\boldsymbol{W}^T = \boldsymbol{Q}$
11: end **for**
12: Updated Codebook **W**

Finally, the updated codebook can be obtained by solving the system of linear equation (14). Therefore, the optimal solutions for encoding the parameters $\boldsymbol{C}$ and the codebook $\boldsymbol{W}$ are obtained. This kind of iterative process is known as coordinate descent method. In this process, we optimize $\boldsymbol{C}(\boldsymbol{W})$ based on the existing value of $\boldsymbol{W}(\boldsymbol{C})$, alternately. The solutions of both $\boldsymbol{C}$ and $\boldsymbol{W}$ are unique, and their sequences also converge to stationary points. In the encoding stage, when codebook $\boldsymbol{W}$ is fixed, we derive the analytical solution of $\boldsymbol{C}$ using (10). Similarly, during the codebook-updating process, the closed form solution of $\boldsymbol{W}$ is derived using (14). The training process is illustrated in Algorithm 1. During the testing stage, each query face image's feature is encoded using the learned codebook $\boldsymbol{W}$. Our experiments also show that, by encoding features at different ages using locality constraint, we can obtain more discriminative feature representation for age-invariant face recognition. The testing process is illustrated in Algorithm 2.

4.3. Important Properties

- In sparse coding, an input feature can be linearly represented in terms of an overcomplete dictionary, which is learned by using a l_1-minimization technique. By incorporating locality information, higher weights are only given

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

to those codebook entries, which are close to the input feature vector. However, this property does not hold for sparse coding.

- In sparse coding, non-zero sparse coefficients can be obtained for more than one class (subject), especially in the presence of noise. This can lead towards the sharing of dissimilar bases for the images of the same subject. However, locality constraint ensures the sharing of similar bases for the images of the same subject.
- It is computationally efficient, as only a few codebook entries (nearest neighbours) are selected for feature encoding, that makes it useful, in practical terms.

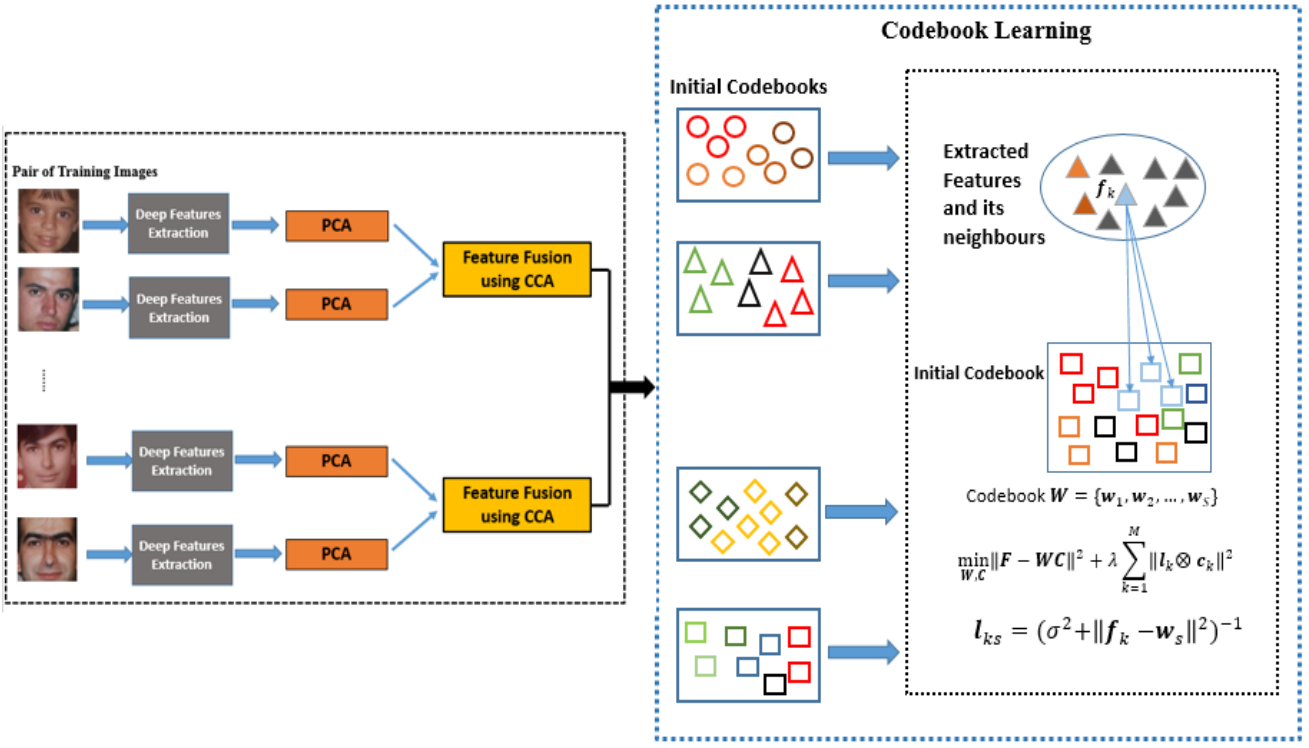


Fig.4. Training Stage of our proposed framework.

5. Feature Fusion using CCA

Codebook learning is the most important and critical step of our proposed algorithm. To recognize faces across different ages, the learned codebook must be discriminative in terms of age progression. To do this, we perform feature-level fusion using CCA. There are different ways to fuse the features. The simplest way is by computing the z-score, which is done by normalizing the two feature vectors and then concatenating them to create a high-dimensional feature vector. The major disadvantage of this method is that it does not take the correlation between the two features into account. In our algorithm, features are extracted and encoded at different ages. Features of the same person at different

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

ages should have a certain degree of correlation. Therefore, we first project the features of the training image's pairs into a coherent feature subspace and then concatenate them to form the final feature vector. This concatenated feature vector is used to learn a discriminative codebook.

Given two sets of training features, with each pair extracted from images of the same subject but at different ages, denoted as $\boldsymbol{F}_{age1}$, and $\boldsymbol{F}_{age2}$, we employ CCA to learn the pairs of directions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which maximize the correlation among the two aging features, such that $\boldsymbol{q}_j^1 = \boldsymbol{\alpha}^T \boldsymbol{F}_{age1}$ and $\boldsymbol{q}_j^2 = \boldsymbol{\beta}^T \boldsymbol{F}_{age2}$, where $\boldsymbol{F}_{age1}$ and $\boldsymbol{F}_{age2}$ are the features extracted from a younger and an older images, respectively. $\boldsymbol{q}_j^1$ and $\boldsymbol{q}_j^2$ are the projected training image's features. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be computed by maximizing the following function.

$$K(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\boldsymbol{\alpha}^T \boldsymbol{C}_{xy} \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^T \boldsymbol{C}_{xx} \boldsymbol{\alpha}. \boldsymbol{\beta}^T \boldsymbol{C}_{yy} \boldsymbol{\beta}}}, \tag{16}$$

where $\boldsymbol{C}_{xx}$ and $\boldsymbol{C}_{yy}$ are the covariance matrices of $\boldsymbol{q}_j{}^1$ and $\boldsymbol{q}_j{}^2$, respectively, while $\boldsymbol{C}_{xy}$ is the cross-variance matrix. After solving (16), $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the eigenvectors of $\boldsymbol{C}_{xx}{}^{-1}\boldsymbol{C}_{xy}\boldsymbol{C}_{yy}{}^{-1}\boldsymbol{C}_{xy}{}^T$, and $\boldsymbol{C}_{yy}{}^{-1}\boldsymbol{C}_{xy}{}^T\boldsymbol{C}_{xx}{}^{-1}\boldsymbol{C}_{xy}$, respectively.

Algorithm 2: Feature encoding using a learned Codebook (Testing Stage)

1: **Input**: $\boldsymbol{W}_s$ for $s \in \{1,2,\ldots,S\}$, $\boldsymbol{q}$ be a feature of a query image.

2: for $k = 1 : N_{knn}$ **do** (N_{knn} is number of nearest neighbors (number of local bases selected from the codebook))

3: The Euclidean locality adaptor $\boldsymbol{l}_k$ is computed

$\boldsymbol{l}_{ks} = (\sigma^2 + \|\boldsymbol{q} - \boldsymbol{w}_{ks}\|^2)^{-1}$

4: Solve Equation (1) to obtain the analytical solution

$\min_{\boldsymbol{c}_k} \|\boldsymbol{q} - \boldsymbol{W}_k \boldsymbol{c}_k\|^2 + \lambda \|\boldsymbol{l}_k \otimes \boldsymbol{c}_k\|_2^2$

$\boldsymbol{\beta}^k = (\boldsymbol{Y}_k + \lambda \text{diag}(\boldsymbol{l}_k)^2)^{-1}\boldsymbol{1}$

$\boldsymbol{c}_k = \boldsymbol{\beta}^k / (1^T \boldsymbol{\beta}^k)$

where $\boldsymbol{Y}_k = (\boldsymbol{q}\boldsymbol{1}^T - \boldsymbol{W}_k)^T (\boldsymbol{q}\boldsymbol{1}^T - \boldsymbol{W}_k)$

5: Similarly, encoding gallery image's features using Step (4). Denote the set of codewords obtained for all the gallery features as $\boldsymbol{G}_k$.

6: Compute the coefficient vector $\boldsymbol{\gamma}_k$ using Equation (17).

7: Compute the residuals $r_k(\boldsymbol{q}) = \|\boldsymbol{c}_k - \boldsymbol{G}_k \boldsymbol{\gamma}_k\|$

8: end **for**

9: identity$(\boldsymbol{q}) = \arg\min_k r_k(\boldsymbol{q})$

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

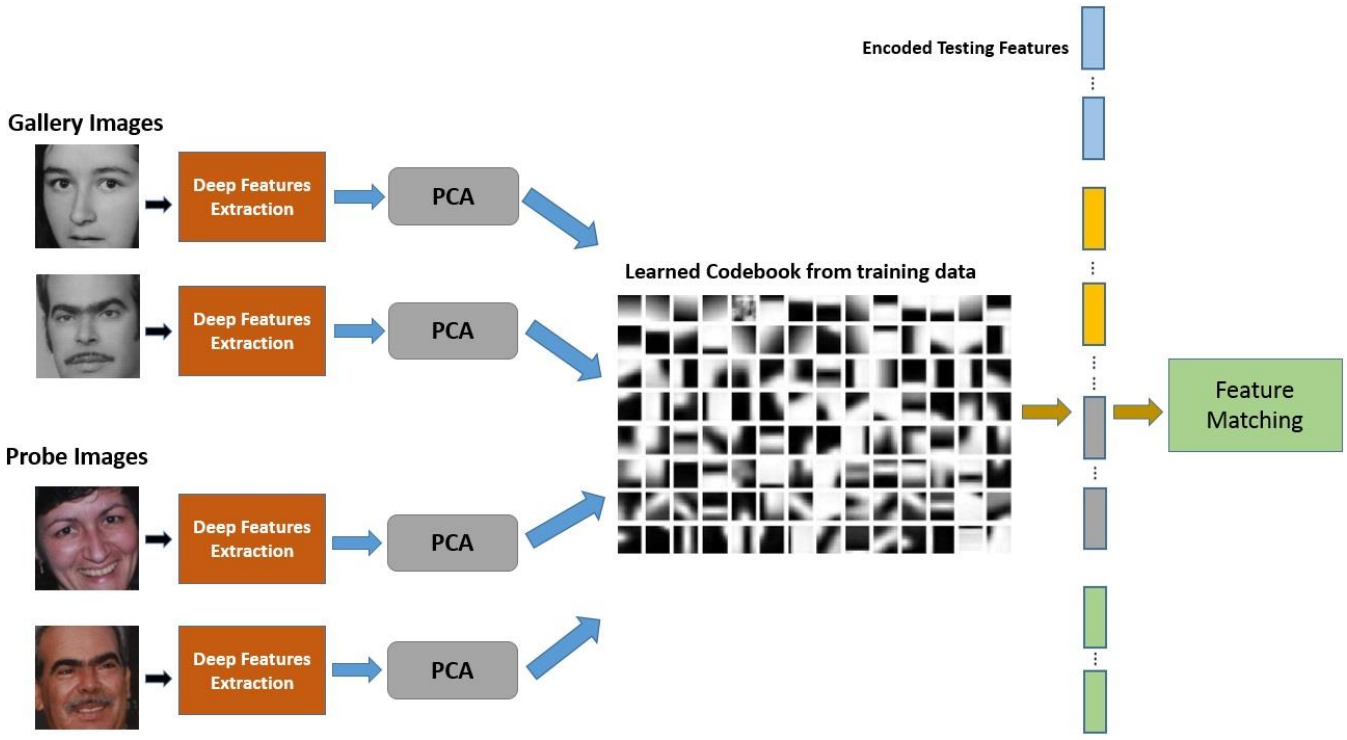


Fig.5. Testing stage of our proposed framework.

6. Feature Matching using Linear Regression

In order to determine the similarity between the encoded gallery and the query images' features, we utilize a linear regression model [13]. The assumption is that the features of different samples lie in a linear subspace, so a query face image can be represented as a linear combination of all its face images in the gallery. The relationship between the query (test) and gallery sample is determined by a coefficient vector $\boldsymbol{\gamma}$. The vector is estimated by using the least-squares method. It is also considered as a prediction problem having a solution based on a regression framework. If a query image $\boldsymbol{q}$ belongs to the k^{th} class, then there must exist a linear relationship between this query image and the gallery samples $\boldsymbol{X}_k$ from the same class, which is defined as follows:

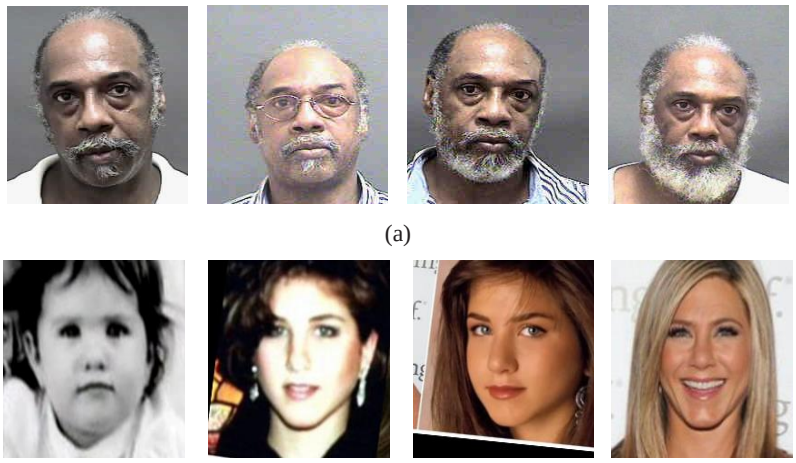
$$\boldsymbol{q} = \boldsymbol{X}_k \boldsymbol{\gamma}_k \,, \tag{17}$$

After estimating the coefficient vector $\boldsymbol{\gamma}$, the corresponding residual values are computed. The decision will be in favor of the gallery image, that has the minimum distance to the query image. Mathematically, it can be written as:

$$j = \min_k \|\boldsymbol{q} - \boldsymbol{X}_k \boldsymbol{\gamma}_k\|. \tag{18}$$

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

Fig.5. shows the flowchart for the testing framework.



(a)

(b)

Fig.6. Sample images with age variations, where each row represents the face images of the same person. (a) MORPH data set, and (b) LAG data set.

7. Experimental Results and Analysis

Creating a data set with a large age variation is a difficult task. Currently, only a few aging datasets are available, which limits the research on age-invariant face recognition. For comprehensive analysis of our proposed AIFR algorithm, face datasets should have the following attributes: (1) many samples per subject, (2) significant age variations among the images of the same person, and (3) images must be taken in unconstrained environments. The performance of our proposed method is evaluated by conducting extensive set of experiments on three challenging face-aging data sets: FGNET [46], MORPH [47], and LAG datasets [48]. There are large variations of pose, expression, and lighting conditions in these data sets. For all the three data sets, we first locate the face region in an image using the Viola-Jones face detector [49], and then resize the face region to 227×227 pixels. In addition to the Linear Regression (LR)-based classifier, we also utilize the Nearest Neighbor (NN) classifier for feature matching. We also evaluate the performance of our proposed feature-encoding framework by fusing two efficient local features, namely Densely sampled Scale Invariant Feature Transform (DSIFT) [50] and Local Binary Pattern Difference Feature (LBPD) [51]. Furthermore, we also evaluate the performance of our proposed method after removing outliers using the Hampel filter [52]. The details regarding the outlier's detection and removal will be explained later in this Section.

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

7.1 Experimental Results on the FGNET Database

FGNET is a challenging data set, containing 1,002 images of 82 subjects taken at different ages, with the minimum age of 0 (less than 12 months), and maximum being 69. Due to the small number of subjects, more images per subject are available. In addition to aging variations, images in this data set also contain large variations in terms of expression, illumination, and pose. For experiments, we follow the same protocol as used in [32, 33], and use the leave-one-out scheme for performance evaluation. We compare our method with various state-of-the-art AIFR methods. These include: (1) the 3D age modeling technique for age estimation and recognition [28]; (2) a discriminative model for AIFR [32]; (3) hidden factor analysis [33], which represents face images with the identity and age components; (4) the maximum entropy feature descriptor [34]; and (5) deep-learning approach based on latent factor guided convolutional neural networks [38]; (6) the method based on the coupled auto-encoder network [40]; and (7) the identity-inference model [37], which has achieved the highest recognition rate so far on this most challenging face aging data set. From Table 1, we can observe that our proposed feature-encoding-based discriminative model clearly outperforms other AIFR methods and achieves the highest rank-1 recognition rate.

TABLE 1

Comparative results in terms of the Rank-1 recognition rate on the FGNET dataset.

Algorithms	Rank-1 Recognition rates
Park et al. [28]	37.4%
Li et al. [32]	47.5%
Gong et al. [33]	69.0%
MEFA [34]	76.2%
CNN-baseline	84.4%
LF-CNN [38]	88.1%
Xu et al. [40]	86.5%
AG-IIM [37]	88.23%
Proposed Method (DSIFT+LBPD) + NN	**90.48%**
Proposed Method (DSIFT+LBPD) + LR	**89.13%**
Proposed Method (Deep features) + NN	**91.46%**
Proposed Method (Deep features) + LR	**90.24%**
Proposed Method (Deep Features) + NN (Hampel)	**92.23%**
Proposed Method (Deep Features) + LR (Hampel)	**91.20%**

7.2 Experimental Results on the MORPH Database

We also conducted experiments on one of the largest publicly available face-aging data set, i.e. MORPH Album 2. This data set contains 78,000 face images from 20,000 subjects. The number of images per subject are small (around

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

4 images per person) due to the availability of large number of subjects. We divide the data set into a training set and a testing set. The training data, used to learn a codebook for feature encoding, contains 20,000 images from 10,000 distinct subjects (2 images per person), with the large age difference. For testing, a gallery set, and a probe set are constructed from the other 10,000 subjects; the probe set consists of 10,000 face images of the oldest age of the subjects, while the gallery set consists of 10,000 face images of the youngest age. The age gap in this data set is 5-6 years. In the experiments, we choose the images with largest age difference between the gallery and probe set. We compared our method with various state-of-the-art conventional and deep-learning-based AIFR methods. Comparative results are tabulated in Table 2, which shows that our method significantly outperforms other state-of-the-art AIFR methods and achieves the highest recognition rate. Sample face images from this data set are shown in Fig.6 (a).

TABLE 2

Comparative results in terms of the Rank-1 recognition rates on the MORPH database (Album 2).

Algorithms	Rank-1 Recognition rates
Park et al. [28]	79.80%
Li et al. [32]	83.90%
Gong et al. [33]	91.14%
MEFA [34]	92.26%
CARC [35]	92.80%
HOG+LPS [36]	94.20%
LPS [36] + Gong et al. [33]	94.87%
LF-CNN [38]	97.51%
AG-IIM [37]	95.62%
AFJT-CNN [39]	97.85%
Li et al. [41]	93.60%
Proposed Method (DSIFT+LBPD) + NN	96.06%
Proposed Method (DSIFT+LBPD) + LR	96.50%
Proposed Method (Deep features) + NN	**97.93%**
Proposed Method (Deep features) + LR	**98.00%**
Proposed Method (Deep Features) + NN (Hampel)	**98.43%**
Proposed Method (Deep Features) + LR (Hampel)	**98.67%**

7.3 Experimental Results on the LAG Database

The Large Age-Gap (LAG) data set [48] was recently released for studying the cross-age face-recognition problem. All the images were taken in the wild, with a very large age difference (0-80) yrs. The data set is created using a Google search. It consists of 3,828 images from 1,010 identities. At least one child and one adult image are available for each identity. Sample face images from this data set are shown in Fig. 6 (b). For performance evaluation, we utilize a two-fold cross validation scheme. Subjects are alternatively assigned to the first and the second fold, and then the average

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

recognition accuracy is computed. The training set is formed by flipping the images in horizontal and vertical directions. The vertically flipped images are further rotated horizontally. There are 8 images of each subject in the training set. The performance of our proposed method is compared with the state-of-the-art high-dimensional LBP feature [53], and some similarity metric learning methods, such as Cosine similarity [54], one shot similarity kernel (OSS) [55], Joint Bayesian [56], and sub-SML [57]. It should be noted that all these similarity metric learning methods used the $fc7$ features extracted from DCNN [9] (trained on the CASIA-Web Face dataset [58]). Results are tabulated in Table 3.

TABLE 3

Comparative results, in terms of the Rank-1 average recognition rates, on the LAG database.

Algorithms	Rank-1 Recognition rates
DCNN [9] + SML [57]	72.43%
DCNN [9] + OSS [55]	66.42%
DCNN [9] + Cosine Similarity [54]	65.08%
DCNN [9] + Joint Bayesian [56]	66.33%
DCNN [9] + CARC [35]	74.82%
HDLBP [53]	71.53%
Bianco et al. [48]	84.95%
Proposed Method (DSIFT+LBPD) + NN	79.88%
Proposed Method (DSIFT+LBPD) + LR	80.00%
Proposed Method (Deep features) + NN	**91.00%**
Proposed Method (Deep features) + LR	**89.44%**
Proposed Method (Deep Features) + NN (Hampel)	**91.63%**
Proposed Method (Deep Features) + LR (Hampel)	**90.50%**

7.4 Parameter Settings

To perform comprehensive analysis of our proposed discriminative model, we evaluate the performance of our algorithm with respect to the number of nearest neighbors selected as the local bases of the codebook, as described in Section 4. As discussed previously, the codebook is first generated by using *k*-means clustering. As the results on the FGNET data set are evaluated using the leave-one-out scheme, we initialize the codebook with $N-1$ entries, where N is the number of subjects in the data set. The computational complexity of the algorithm depends on the number of nearest neighbors of the feature descriptor $\boldsymbol{f}_k$, which can also be considered as the local bases $\boldsymbol{w}_k$. A small number of neighbors will lead to faster computation.

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

To search the nearest neighbors, the kNN search approach based on the hierarchical model [59] is utilized. The approach quantizes each descriptor into P subspaces. A codebook is then applied for each subspace. For the FGNET data set, we measured the recognition rates with different numbers of nearest neighbors, from 50 to 150, and the results are shown in Fig. 10(b). When we increase the number of neighbors, the recognition rate also increases. With the FG-NET data set, the highest recognition rate obtained is 91.46%, by searching for 150 nearest neighbors to form the local bases in the computation of the codes. For the MORPH data set, recognition results are recorded with different numbers of nearest neighbors, ranging from 20 to 150, as shown in Fig. 11(b). The highest recognition rate obtained is 98.00%, which is slightly better than the deep-learning-based method [39]. It is found that the modeling capacity can be greatly improved by using a larger codebook. For the LAG data set, the highest recognition rate of 89.44% is achieved, using the linear-regression (LR)-based classifier. The parameter σ in Equation (2) is set to 0.5. The value of σ controls the locality error of the encoding scheme. Theorem 2 in [60] shows that locality error reduces, as the value of σ decreases. Therefore, the value of σ must be as small as possible. In our experiments, we vary the value of σ from 0.1 to 0.7 and found that $\sigma = 0.5$ gives the best recognition performance. The choice of the parameter λ in Equation (1) can be well explained by Equation (10). It is worth noting that the matrix $\boldsymbol{Y}$ is symmetric as well as semi-positive. If $\boldsymbol{Y}$ becomes singular or close to singular, the matrix $\boldsymbol{Y} + \lambda \mathrm{diag}(\boldsymbol{l}_k)^2$ is still conditioned. The reason for this is that $\lambda \mathrm{diag}(\boldsymbol{l}_k)^2$ penalizes large distances and captures the correlation among the data samples. By choosing $\lambda < 10^{-6}$, the matrix becomes singular or close to singular, which will produce inaccurate results. Experiments show that $\lambda = 0.001$ provides the optimal recognition results. Moreover, the performance of our method is also evaluated on all the three data sets, using deep features of different dimensions, and results are tabulated in Figs. 10, 11, and 12(a), respectively. In addition to this, we also evaluated the face-recognition performance, with and without performing the proposed feature-encoding scheme. The results using the nearest neighbor (NN) classifier and the linear regression (LR) classifier are shown in Fig. 7. It can be observed that our proposed feature-encoding framework boosts the recognition accuracy by 20-35%.

As discussed before, CCA enhances the correlation among the images of the same subject taken at different ages. To evaluate the superiority of CCA for feature fusion, we perform the comparative analysis of CCA-based feature fusion with simple concatenation of the two features. Experimental results are reported for all the three datasets with

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

optimal feature dimensions and the number of nearest neighbors used for feature encoding. Fig. 8 shows the corresponding comparative analysis. It can be observed that the feature-fusion framework based on CCA brings a significant improvement in recognition rate.

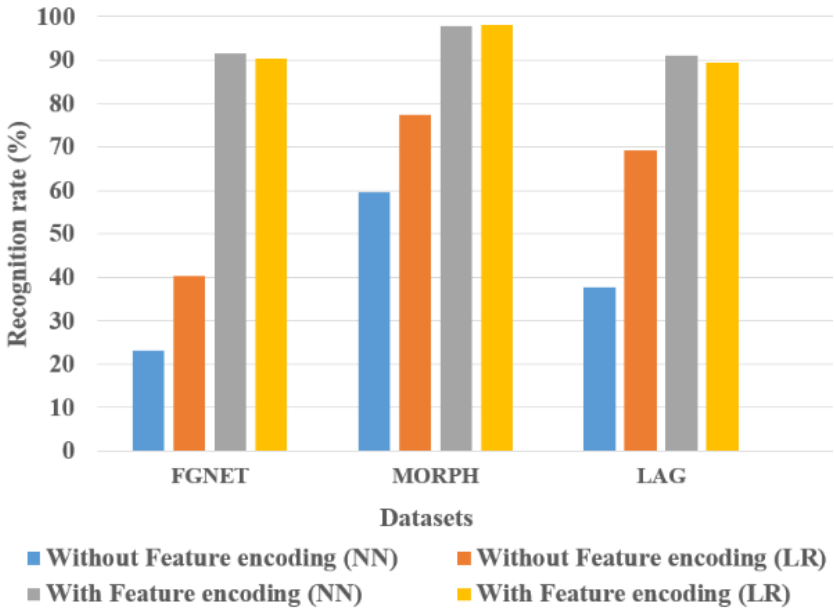


Fig.7. Recognition rates with and without performing feature encoding for all the three data sets at the corresponding optimal feature dimensions.

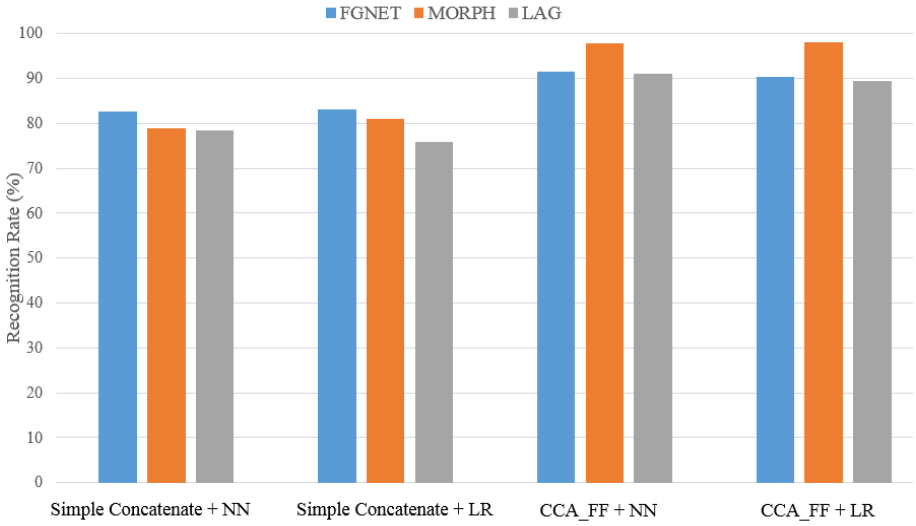


Fig.8. Recognition rates with and without performing feature fusion using CCA for all the three data sets at the corresponding optimal feature dimensions.

Another important parameter of our proposed framework is the number of image pairs per person, which are used for the CCA pairwise training. For the FGNET data set, 10 images per subject are available. Therefore, we divide these

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

images into two subsets, with 5 images each, and then use them for CCA training. One subset consists of those younger images, while the other one contains the older images. As the performance on the FGNET data set was evaluated using the leave-one-out cross validation scheme, we used 81 image pairs for training. For the MORPH data set, 10,000 image pairs were used. The training set consists of 10,000 subjects, with the two images having the largest age difference for each subject. Images of each subject were divided equally into two subsets for pairwise CCA training. For the LAG data set, the two-fold scheme was used for performance evaluation. For training, we selected eight images per subject, with a large age difference. As the total number of subjects in this dataset was 1,010, we used 505 image pairs for CCA training in each fold.

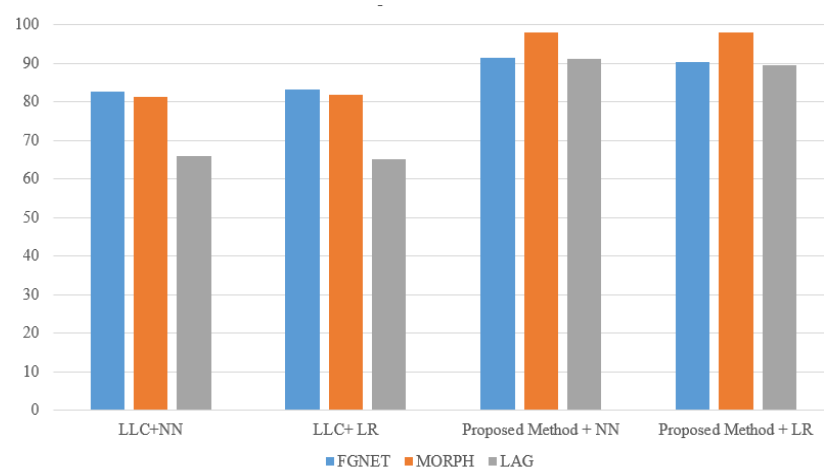


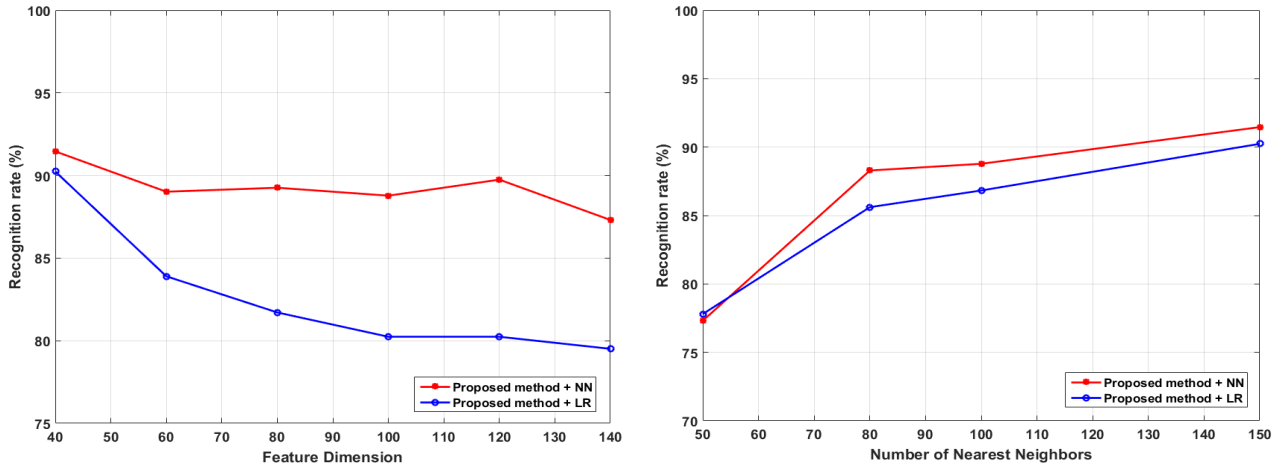
Fig.9. Comparative analysis with respect to LLC [8] at the optimal feature dimensions and number of nearest neighbors.

7.5. Difference from LLC

In this section, we will explain the major differences between the proposed feature encoding scheme and LLC [8]. LLC provides an approximated solution by using different constraints. One of the constraints used by LLC is the norm-bounded constraint $\|\boldsymbol{w}_s\|_2 \leq 1$, where each codebook entry is required to have a unit length. However, we found that this constraint is not essential in the codebook learning process, due to two main reasons. First, due to the locality constraint, the length of columns of $\boldsymbol{W}$ cannot be large. Second, better optimized values can be obtained for the proposed objective function, due to the smaller number of constraints. By removing this constraint, we can obtain a better optimized codebook that better incorporates the local structure of the data samples. This results in an improved classification performance. Moreover, we can also obtain the closed-form solutions for both the sparse coding phase and the codebook updating stage. Our proposed algorithm directly minimizes $\sum_{k=1}^{M} f(\boldsymbol{W}, \boldsymbol{f}_k)$, while LLC [8]

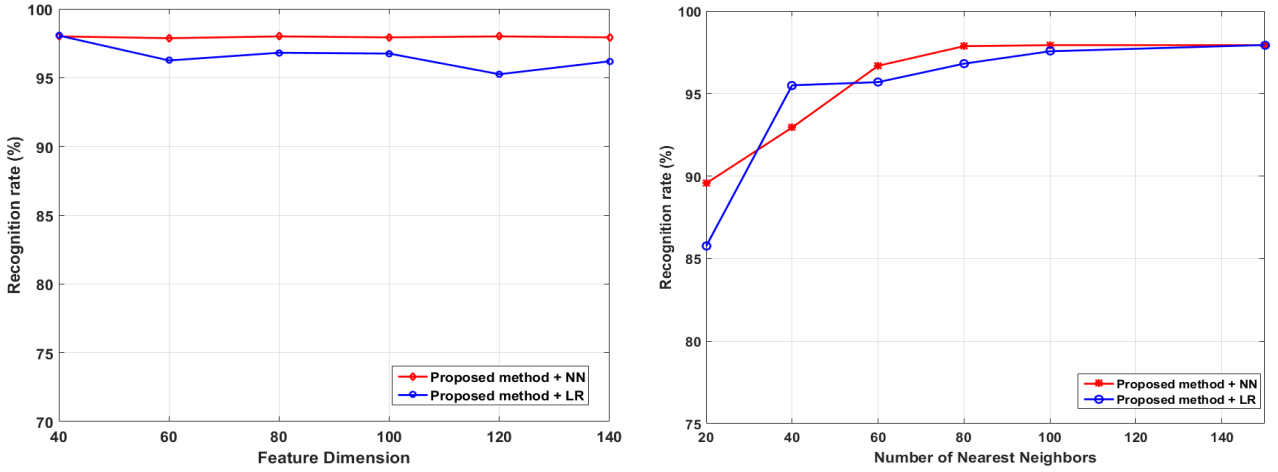
* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)

minimizes $f(\boldsymbol{W}, \boldsymbol{f}_k)$, when the feature $\boldsymbol{f}_k$ is drawn from $\boldsymbol{F}$, and hence provides an approximated form of the objective function. Another major difference is that LLC does not incorporate the locality information during the process of codebook optimization. However, our method uses the locality information during both the codebook-updating and encoding stages. Furthermore, we compare our proposed method with LLC in terms of recognition rate. To obtain the optimal performance from LLC, we keep the parameters the same as those described in the original LLC paper. After the process of deep feature extraction and fusion, we employ the LLC algorithm to encode the features. The comparative results are shown in Fig. 9. It can be observed that our proposed feature encoding scheme outperforms LLC in terms of recognition rates and achieves superior performance.



(a) (b)

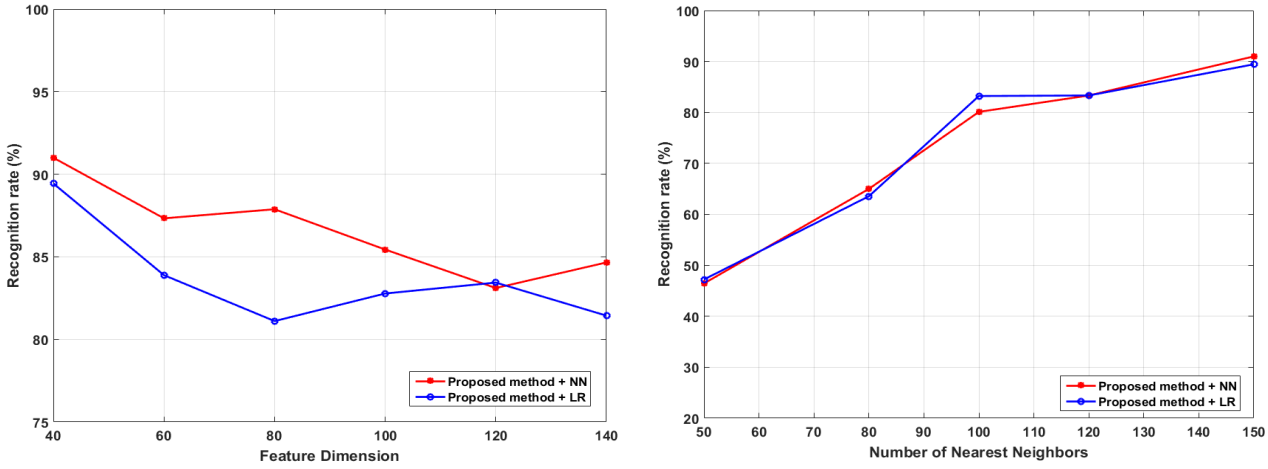
Fig.10. Recognition rates obtained on the FGNET data set. (a) Feature dimensions with 150-NN, and (b) Number of nearest neighbors (40-D features).



(a) (b)

Fig.11. Recognition rates obtained on the MORPH data set. (a) Feature dimensions with 150-NN, and (b) Number of nearest neighbors (40-D features).

* Corresponding author
Email addresses: 15902620r@connect.polyu.hk (M. Saad Shakeel)
enkmlam@polyu.edu.hk (K.-M. Lam)



(a) (b)

Fig.12. Recognition rates obtained on the LAG data set. (a) Feature dimensions with 150-NN, and (b) Number of nearest neighbors (40-D features).

7.6. Outlier Detection and Removal

In the feature matching stage, there may exist some outliers in the gallery set, which can create problems in learning the relationship between the gallery and the probe images. Therefore, it is better to first detect and remove the outliers from the gallery set. In order to detect and remove the outliers, we utilize the Hampel filter [52]. In this regard, an image is first transformed into the frequency domain using Fast Fourier Transform (FFT), and then Hampel identifier is used to detect outliers in both the real and imaginary spectra. The detected outliers are replaced by the interpolated values, and then the clean spectrum is transformed into the original domain using inverse Fourier transform. The outliers are identified by first computing the median of a window consisting of the sample and its six neighboring samples, three on each side. In the second step, the standard deviation is computed with respect to the median of the window using the Median Absolute Deviation (MAD). If the difference between the sample and the mean exceeds three times the standard deviation, then it will be replaced by the median. The only tuning parameter in this filtering operation is the m neighbors in the measurement window. In our experiments, we set $m = 4$, which provides the optimal performance. The number of outliers identified in the FGNET, MORPH, and LAG dataset are 0.36%, 4.18%, and 0.44%, respectively. In this paper, those outliers are denoted as internal outliers. After removing these outliers using the Hampel filter, 0.2-0.6% improvement in recognition rate is observed. To further evaluate the robustness of our proposed method, we manually add outliers to the gallery set, which are called external outliers. The outliers are samples, which are a certain multiple of the standard deviation from the mean. Mathematically, it can be written as $mean\,(Gallery) + (k \times Std)$. In our experiments, we set $k = 4$. The number of outliers in FGNET, MORPH, and

LAG are increased to 4.27%, 9.02%, and 1.25%, respectively. After adding the outliers, we first compute the recognition rate without using the Hampel filter. We observed that our proposed method shows high robustness even against the large number of outliers. For all the three datasets, only 4-5% decline in recognition rate is observed. In the second phase, we apply the Hampel filter for removing the added outliers, which also provides promising results. Fig. 13 shows the recognition rates of our proposed method, with and without using the Hampel filter for both internal and externally added outliers. The recognition results are reported at the optimal feature dimensions and the optimal number of nearest neighbors used for encoding.

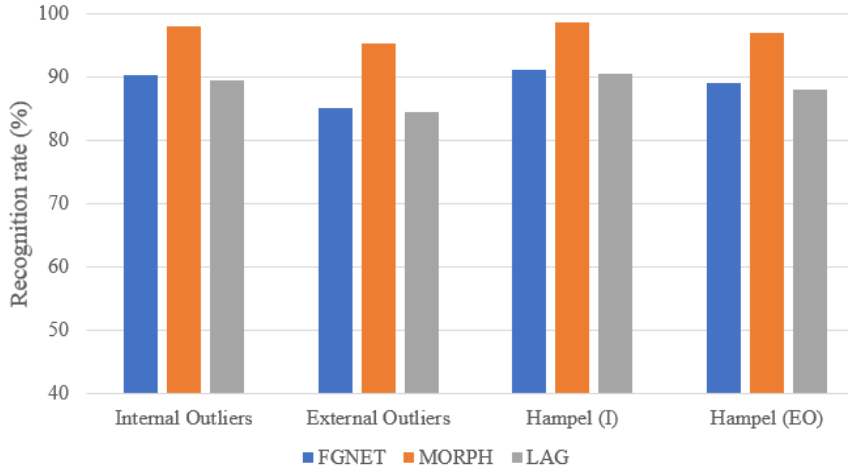
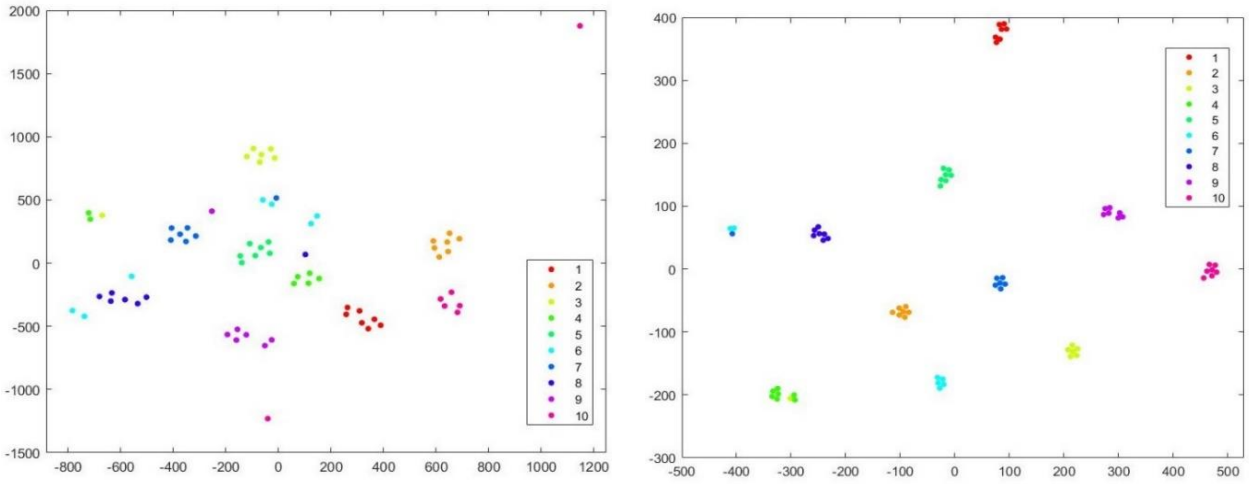


Fig.13. Recognition rates with and without using the Hampel filter, in the presence of internal, as well as externally added outliers, where Hampel (I) and Hampel (EO) represent the outlier removal operation for internal and external outliers, respectively.

7.7 Overall Benchmark Comparison

In our experiments, the performance of our proposed method is compared with state-of-the-art aging face recognition methods. For the MORPH Album 2 data set, all the methods compared have the same settings as those used in their original literature, and the same protocol is used for the training set and the testing set. 10,000 pairs of face images are used for training, while the remaining subjects are used for testing. FGNET is one of the most challenging data sets for aging-face recognition. All the images were taken in the wild, with the age gap of 0-45 years. Our proposed method achieves the highest rank-1 recognition rate on this data set, which is 91.46% by using deep features. Recognition results were recorded using the NN classifier and the LR-based classifier. It is worth noting that in our proposed algorithm, the closed-form solution for $\boldsymbol{W}$ can be derived using (14), in which the matrix $\boldsymbol{P}$ is positive definite. On the other hand, analytical solution of $\boldsymbol{C}$ can be obtained as shown in (10). Our proposed algorithm is also computationally efficient as it does not need to solve the l_1 minimization problem.



(a) (b)

Fig. 14. Visualization of the learned features before and after encoding using t-SNE. (a) before encoding and (b) after encoding.

7.8 Better Reconstruction

According to the sparse-coding theory, each feature can be represented as a sparse linear combination of codewords for reconstruction. However, it is sensitive to feature variance, which may lead to selecting codewords far away from the input test feature in reconstruction. In this way, two similar features may select different codewords. Conversely, our method is based on locality information in searching the nearest neighbors, so only the codebook entries near to the input (test) feature are selected, and the selected local visual codewords can achieve better feature reconstruction. In other words, locality constraint ensures that the features of the same person at different ages have the same codewords. Learning the codebook with the locality adaptor offers the following advantages: (1) non-smooth optimization can be avoided due to the use of a smooth objective function, and (2) codebook size with locality adaptor has nothing to do with the data's dimension. For better understanding of our proposed feature encoding framework, we visualize the deep features, obtained before and after encoding, using t-distributed Stochastic Neighbor Embedding (t-SNE) [61]. For visualization, we randomly selected 10 subjects, each with 7 images from the LAG dataset. Fig. 14 shows the deep features learned before and after applying our encoding framework. The encoded features are well separated in the feature subspace. In other words, features of the same identity are represented by the same or similar codewords.

7.9 Robustness to Noise Variations

In face recognition, a query face image may suffer from noise, due to various factors, such as environmental conditions, transmission errors, etc. These noise variations degrade the performance of face recognition systems. In

this section, we evaluate the performance of our proposed method with respect to noise variations. Previously proposed methods [33, 34, 40] first decompose a face image into identity, aging, and noise components, and then use the identity component for recognition. Xu et al. [40] proposed a coupled autoencoder to eliminate the noisy component from input images, and then performed recognition. It is worth noting that these methods only consider the inherent noise in images, not the externally added noises, such as Gaussian noise, salt & pepper, etc.

(a) (b) (c)

Fig. 15. Original images and noisy images obtained after adding Gaussian noise. (a) FGNET, (b) MORPH, and (c) LAG data sets.

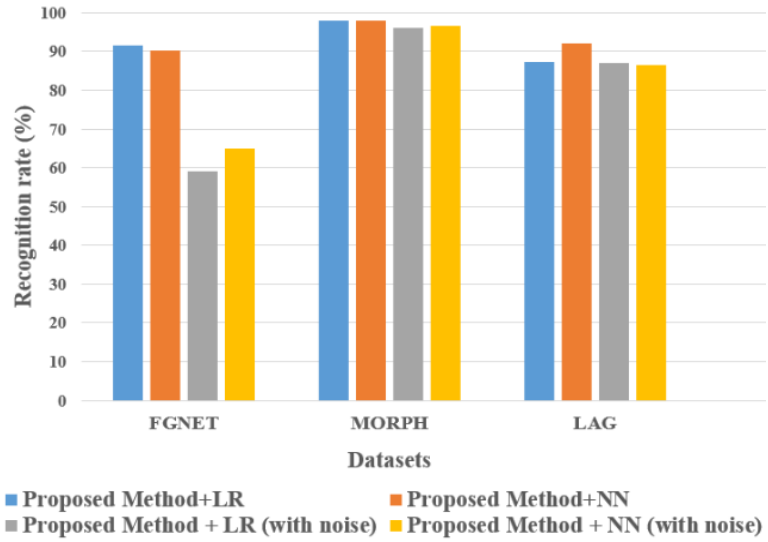


Fig. 16. Recognition rates of our proposed method on all the three data sets, with and without noise variations.

To evaluate the robustness of our algorithm against noise variations, we add external Gaussian noise into the probe face images. The added Gaussian noise affects 30% of the pixels in each probe image. Our experimental results show the robustness of our proposed method in the presence of noise variations. Only a 3-4% drop in recognition performance is observed for the MORPH and LAG data sets. However, a 20-25% drop in recognition rate is observed for the FGNET data set. The reason for this is that some images in the FGNET data set are of very poor quality, especially those childhood images. On the other hand, face images in the MORPH and LAG data sets have better visual quality, as compared to the FGNET data set. This problem can be solved by using some denoising filters, prior to feature extraction. Images obtained, after adding external Gaussian noise, are shown in Fig. 15. Fig. 16 shows the recognition results of our proposed method, with and without noise variations, using the NN and the LR-based classifiers [13].

7.10 Computational Complexity Analysis

Our proposed method consists of three major parts, which are: (1) feature fusion using CCA, (2) codebook learning, and (3) feature encoding. In this section, we will evaluate the computational time of these three learning stages. As discussed before, the computational complexity of the feature-encoding stage depends on the number of nearest neighbors selected for encoding. Therefore, we report the results using the corresponding optimal feature dimension and the number of nearest neighbors for all the three datasets. Table 4 shows the computation times in seconds for both the feature-fusion and the codebook-learning processes. The computational complexity with the FGNET dataset is much lower than the MORPH and LAG data sets, as the size of the training set is small. The training set based on the MORPH and LAG data sets consist of 20,000 and 4,000 images, respectively. For the MORPH data set, the results are reported with a feature dimension of 40, and 150-NN. For the LAG data set, the results are reported with a feature dimension of 40, and 100-NN. Table 5 tabulates the computation times in seconds of performing feature encoding for the whole testing set, and a single testing image.

TABLE 4

Computation time in seconds for the two stages of learning (Training).

Data set	Feature Fusion using CCA	Codebook Learning
FGNET	0.011s	0.88s
MORPH	0.264s	54.80s
LAG	0.120s	16.10s

TABLE 5

Computation time in seconds of performing feature encoding for all the three datasets (Testing).

Data set	Feature encoding (single image)	Feature encoding (Whole data set)
FGNET	0.0029s	0.12s
MORPH	0.0096s	9.60s
LAG	0.0052s	0.26s

7.11 Comparison with Local Feature Descriptors

We also evaluate our proposed feature-encoding framework using two local feature descriptors, namely DSIFT [50] and LBPD [51]. To extract local features, each face image is divided into non-overlapping patches, and the selected feature descriptors are extracted from each of the patches. The extracted features are then concatenated to create a high-

dimensional feature vector. In our proposed method, we perform dense sampling of the SIFT feature descriptors from the whole face image, which is equivalent to placing a regular grid on a face region as shown in Fig. 17.

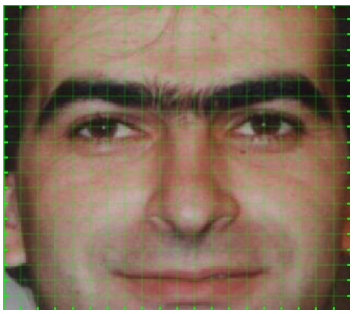


Fig.17. Placement of a regular grid on a face image using DSIFT.

Better recognition performance can be achieved by extracting information from dense grids, instead of a few sparse key points. In this way, we can extract the information about the distribution of edge directions in the entire face region, which has been proved to be age-invariant discriminant information in [32].

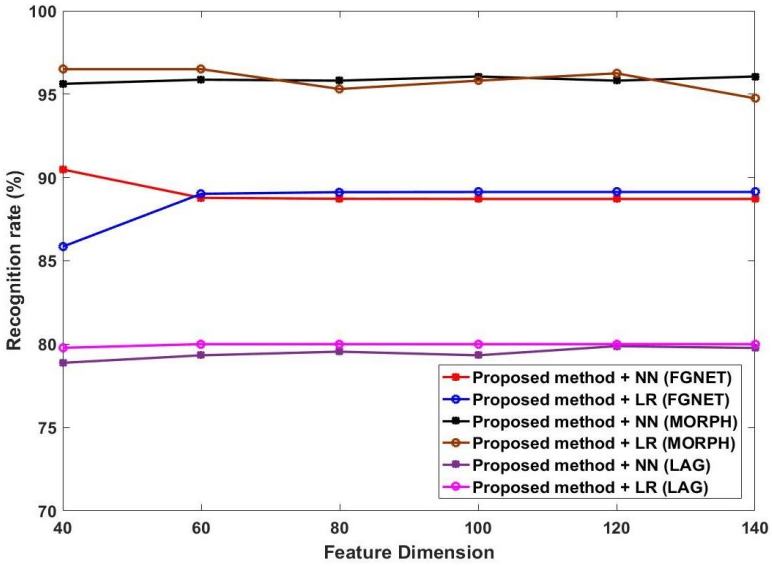


Fig. 18. Recognition rates with different feature dimensions using local feature descriptors (DSIFT+LBPD), with 150 nearest neighbors.

Recently, a numerical variant of LBP [51], known as LBPD, was proposed, which offers several advantages over LBP. The LBPD feature is computed by taking the difference between the LBP codes and its corresponding mean of a given local region. After extracting these two local features, we concatenate them to form a final feature vector, which is of very high dimensionality. To reduce the feature dimension, PCA is employed, before performing our proposed feature-encoding method. Recognition rates with different feature dimensions for all the three data sets are shown in Fig. 18. It is observed that our method can obtain state-of-the-art performance with local feature descriptors but there

are some drawbacks. First, the DSIFT feature is computationally expensive, as it requires computing the gradient for each pixel. Second, the handcrafted features are sensitive to noise, and their performance heavily depend on the preprocessing of face images.

8. Conclusion

In this paper, we have proposed a robust feature-encoding-based discriminative model for age-invariant face recognition. Our method extracts high-level deep features using a pre-trained Deep-CNN model (AlexNet), and then performs feature encoding by using a Euclidean locality adaptor. By using the locality information, we can obtain closed form solutions for both the sparse-coding phase and the codebook updating stage. It can be observed that the utilization of locality information during the codebook-updating process better incorporates the local structure of the data samples, and provides improved classification performance. We also show that the utilization of the norm-bounded constraint $\|\boldsymbol{w}_s\|_2 \leq 1$ is not essential for the codebook learning process. By using a lesser number of constraints, better optimized values can be obtained. Furthermore, using CCA for feature fusion takes into consideration the correlation among the features of the same person at different ages. This helps in learning a codebook insensitive to aging variations. At the testing stage, the query and gallery image's features are encoded using a learned codebook, with very low computational complexity. Our proposed method encodes the identity of the same person at different ages using the same or similar codewords. Experiment results on three challenging face-aging data sets show that our method outperforms the state-of-the-art age-invariant face recognition methods, and obtains the highest recognition accuracy. In our future work, we will focus on designing a new convolutional layer, which can be embedded into the feature extraction network to realize both feature fusion and codebook learning in an end-to-end manner. Therefore, the optimization of feature representation and feature fusion can be achieved at the same time and the correlated complementary information between the two steps can be maximized.

Acknowledgment

The work described in this paper was supported by the GRF Grant PolyU 152765/16E (project code: B-Q55J) of the Hong Kong SAR Government.

References

[1] M. Bereta, P. Karczmarek, W. Pedrycz, M. Reformat, Local descriptors in application to the aging problem in face recognition, Pattern Recognit. 46 (10) (2013) 2634-2646.

[2] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of key points, in: Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision, 2004, pp. 1-22.

[3] A. Bosch, A. Zisserman, X. Munoz, Scene classification using a hybrid generative/ discriminative approach, IEEE Trans. Pattern Anal. Mach. Intell. 30(4) (2008) 712-727.

[4] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 2169-2178.

[5] L. Yang, R. Jin, R. Sukthankar, F. Jurie, Unifying discriminative visual codebook generation with classifier training for object category recognition," in in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-8.

[6] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1794-1801.

[7] K. Yu, T. Zhang, Y. Gong, Nonlinear learning using local coordinate coding, in: Neural Information Processing Systems (NIPS), 2009, pp. 2223-2231.

[8] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3360-3367.

[9] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891-1898.

[10] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: A unified embedding for face recognition and clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815-823.

[11] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep Face Recognition, in: British Machine Vision Conference (BMVC), 2015, pp. 6.

[12] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, SphereFace: Deep Hypersphere Embedding for Face Recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[13] I. Naseem, R. Togneri, M. Bennamoun, Linear regression for face recognition, IEEE Trans. Pattern Anal. Mach. Intell, 32 (11) (2010) 2106–2112.

[14] Y. Fu, T. S. Huang, Human age estimation with regression on discriminative aging manifold, IEEE Trans. Multimedia, 10 (4) (2008) 578–584.

[15] G. Guo, Y. Fu, C. R. Dyer, T. S. Huang, Image-based human age estimation by manifold learning and locally adjusted robust regression, IEEE Trans. Image Process., 17 (7) (2008) 1178–1188.

[16] N. Ramanathan, R. Chellappa, Face verification across age progression, IEEE Trans. Image Process., 15 (11) (2006) 3349–3361.

[17] W-L. Chao, J-Z. Liu, J-J. Ding, Facial age estimation based on label-sensitive learning and age-oriented regression, Pattern Recognit., 46 (3) (2013) 628-641.

[18] H. Dibeklioglu, F. Alnajar, A. Salah, T. Gevers, Combining facial dynamics with appearance for age estimation, IEEE Trans. Image Process., 24 (6) (2015) 1928-1943.

[19] J K. Pontes, A S. Britto Jr, C. Fookes, A L. Koerich, A flexible hierarchical approach for facial age estimation based on multiple features, Pattern Recognit., 54 (C) (2016) 34-51.

[20] M.T. B. Iqbal, M. Shoyaib, B. Ryu, M. A. Wadud, O. Chae, Directional Age-Primitive Pattern (DAPP) for human age group recognition and age estimation, IEEE Trans. Inf. Forens. Security, 12 (11) (2017) 2505-2517.

[21] H. Liu, J. Lu, J. Feng, J. Zhou, Group-aware deep feature learning for facial age estimation, Pattern Recognit., 66 (2017) 82-94.

[22] H. Liu, J. Lu, J. Feng, and J. Zhou, Ordinal deep learning for facial age estimation, IEEE Trans. Circuit. Sys. Video. Tech, (2017) DOI:10.1109/TCSVT.2017.2782709.

[23] J.-X. Du, C.-M. Zhai, Y.-Q. Ye, Face aging simulation and recognition based on NMF algorithm with sparseness constraints, Neurocomputing, (116) (2013) 250–259.

[24] H. Yang, D. Huang, Y. Wang, H. Wang, Y. Tang, Face aging effect simulation using hidden factor analysis joint sparse representation, IEEE Trans. Image Process., 25 (6) (2016) 2493-2507.

[25] Y. Li, Y. Li, Face aging effect simulation model based on multilayer representation and shearlet transform, Journal of Electronic Imaging, 26 (5) (2017).

[26] S E. Choi, J. Jo, S. Lee, H. Choi, I-J Kim, J. Kim, Age face simulation using aging functions on global and local features with residual images, Expert systems with applications, 80 (2017) 107-125.

[27] A. Lanitis, C. J. Taylor, T. F. Cootes. Toward automatic simulation of aging effects on face images. IEEE Trans. Pattern Anal. Mach. Intell., 24(4) (2002) 442–455.

[28] U. Park, Y. Tong, A. K. Jain, Age-invariant face recognition, IEEE Trans. Pattern Anal. Mach. Intell., 32 (5) (2010) 947–954.

[29] J. Suo, X. Chen, S. Shan, W. Gao, Learning long term face aging patterns from partially dense aging databases, in IEEE Conference on Computer Vision (ICCV), 2009, pp. 622–629.

[30] J. Suo, S.-C. Zhu, S. Shan, X. Chen, A compositional and dynamic model for face aging, IEEE Trans. Pattern Anal. Mach. Intell., 32 (3) (2010) 385–401.

[31] H. Ling, S. Soatto, N. Ramanathan, D. W. Jacobs, Face verification across age progression using discriminative methods, IEEE Trans. Inf. Forens. Security, 5 (1) (2010) 82–91.

[32] Z. Li, U. Park, A. K. Jain, A discriminative model for age invariant face recognition, IEEE Trans. Inf. Forens. Security, 6 (2) (2011) 1028–1037.

[33] D. Gong, Z. Li, D. Lin, J. Liu, X. Tang, Hidden factor analysis for age invariant face recognition, in: IEEE Conference on Computer Vision (ICCV), 2013, pp. 2872–2879.

[34] D. Gong, Z. Li, D. Tao, A Maximum Entropy Feature Descriptor for age-invariant face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5289-5297.

[35] B.-C. Chen, C.-S. Chen, W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in European Conference on Computer Vision (ECCV), 2014, pp. 768–783.

[36] Z. Li, D. Gong, X. Li, D. Tao, Aging face recognition: A Hierarchical learning model based on local patterns selection, IEEE Trans. Image Process, 25 (5) (2016) 2146-2154.

[37] H. Zhou, K-M Lam, Age-invariant face recognition based on identity inference from appearance age, Pattern Recognit. 76 (2018) 191-202.

[38] Y. Wen, Z. Li, Y. Qiao, Latent Factor Guided Convolutional Neural Networks for Age-Invariant Face Recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4893-4901.

[39] H. Li, H. Hu, C. Yip, Age-Related Factor Guided Joint Task Modelling Convolutional Neural Network for Cross-Age Face Recognition, IEEE Trans. Inf. Forens. Security, 13 (9) (2018) 2383-2392.

[40] C. Xu, Q. Liu, M. Ye, Age invariant face recognition and retrieval by coupled auto-encoder networks, Neurocomputing, 222 (2017) 62-71.

[41] Y. Li, G. Wang, L. Nie, Q. Wang, W. Tan, Distance metric optimization driven convolutional neural network for age invariant face recognition, Pattern Recognit. 75 (2018) 51-62.

[42] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: Neural Information Processing Systems (NIPS), 2012, pp. 1106-1114.

[43] A. Elgammal, R. Duraiswami, L. Davis, Efficient kernel density estimation using the fast gauss transform with applications to color modeling and tracking, IEEE Trans. Pattern Anal. Mach. Intell., 25 (11) (2003) 1499–1504.

[44] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science, 290 (5500) (2000) 2323-2326.

[45] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, D. Cai, Graph regularized sparse coding for image representation, IEEE Trans. Image Process., 20 (5) (2011) 1327-1336.

[46] Facial Image Processing and Analysis (FIPA). FG-NET Aging Database. [Online]. Available: http://fipa.cs.kit.edu/433.php#Downloads.

[47] K. Ricanek, Jr., T. Tesafaye, MORPH: A longitudinal image database of normal adult age-progression, in: IEEE Conf. Automat. Face Gesture Recog., 2006, pp. 341–345.

[48] S. Bianco, Large Age-gap Face verification by Feature Injection in Deep Networks, Pattern Recognit. lett, 90 (2017) 36-42.

[49] P. Viola, M. J. Jones, Robust real-time face detection, International Journal of Computer Vision, 57 (2) (2004) 137-154.

[50] A. Vedaldi, B. Fulkerson, Vlfeat: An open and portable library of computer vision algorithms," in: International Conference on Multimedia, 2010, pp. 1469-1472. Available: www.vlfeat.org/

[51] X. Hong, G. Zhao, M. Pietikainen, X. Chen, Combining LBP Difference and Feature correlation for texture description, IEEE Trans. Image Process, 23 (6), 2014.

[52] D-P. Allen, A frequency domain Hampel filter for blind rejection of sinusoidal interference from electromyograms, Journal of Neuroscience Methods, 177 (2) (2009) 303-310.

[53] D. Chen, X. Cao, F. Wen, J. Sun, Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3025-3032.

[54] H.V. Nguyen, L. Bai, Cosine similarity metric learning for face verification, in: Asian Conference on Computer Vision (ACCV), 2010, pp. 709-720.

[55] L. Wolf, T. Hassner, Y. Taigman, The one-shot similarity kernel, in: IEEE Conference on Computer Vision (ICCV), 2009, pp. 897–902.

[56] D. Chen, X. Cao, L. Wang, F. Wen, J. Sun, Bayesian face revisited: A joint formulation, in: European Conference on Computer Vision (ECCV), 2012, pp. 566–579.

[57] Q. Cao, Y. Ying, P. Li, Similarity metric learning for face recognition, in: IEEE Conference on Computer Vision (ICCV), 2013, pp. 2408–2415.

[58] D. Yi, Z. Lei, S. Liao, S. Z. Li, Learning face representation from scratch, arXiv preprint arXiv: 1411.7923, 2014.

[59] J. L. Bentley, Multidimensional Binary Search Trees Used for Associative Searching, Comm. ACM, 18 (1975) 509-517.

[60] J. Pang, L. Qin, C. Zhang, W. Zhang, Q. Huang, B. Yin, Local Laplacian coding from Theoretical Analysis of Local Coding Schemes for Locally Linear Classification, IEEE Trans. Cybern., 45 (12) (2015) 2937-2947.

[61] V.D. Maaten, "Visualizing data using t-SNE," *J. Mach. Learn. Res*., vol. 9, pp. 2431-2456, 2008.

[62] X. Han, Y. Zhong, L. Cao, L. Zhang, Pre-trained Alex Net Architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification, Remote sensing, 9 (8) (2017), pp. 848.

M. Saad Shakeel received the B.Eng. degree in electrical engineering from the University of the Punjab, Pakistan, and M. Eng. Degree in electrical & computer engineering from the South China University of Technology, Guangzhou, China. He is currently pursuing the Ph.D. degree with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong. His research interests include image processing, computer vision, and pattern recognition.

Kin-Man Lam (SM'14) received the Associateship in electronic engineering from The Hong Kong Polytechnic University, Hong Kong, in 1986; the M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College of Science, Technology and Medicine, London, U.K., in 1987; and the Ph.D. degree from the Department of Electrical Engineering, University of Sydney, Sydney, NSW, Australia, in 1996. He is currently a Professor with The Hong Kong Polytechnic University. His research interests include human face recognition, image and video processing, and computer vision.